Histograms

What is a histogram?

A histogram is one of the basic quality tools. It is used to graphically summarize and display the distribution and variation of a process data set. A frequency distribution shows how often each different value in a set of data occurs. The main purpose of a histogram is to clarify the presentation of data. You can present the same information in a table; however, the graphic presentation format usually makes it easier to see relationships. It is a useful tool for breaking out process data into regions or bins for determining frequencies of certain events or categories of data. These charts can help show the most frequent.

Typical applications of histograms in root cause analysis include:

- Presenting data to determine which causes dominate
- Understanding the distribution of occurrences of different problems, causes, consequences, etc.

A histogram can typically help you answer the following questions:

- What is the most common system response?
- What distribution (center, variation and shape) does the data have?
- Does the data look symmetric or is it skewed to the left or right?

A histogram is a specialized type of bar chart. Individual data points are grouped together in classes, so that you can get an idea of how frequently data in each class occur in the data set. High bars indicate more points in a class, and low bars indicate less points.

Weaknesses

There are two weaknesses of histograms that you should bear in mind. The first is that histograms can be manipulated to show different pictures. If too few or too many bars are used, the histogram can be misleading. This is an area which requires some judgment, and perhaps some experimentation, based on the analyst's experience.

Histograms can also obscure the time differences among data sets. For example, if you looked at data for #births/day in the United States in 1996, you would miss any seasonal variations, e.g. peaks around the times of full moons. Likewise, in quality control, a histogram of a process run tells only one part of a long story. There is a need to keep reviewing the histograms and control charts for consecutive process runs over an extended time to gain useful knowledge about a process.

Histogram statistics[1]:

For histograms, the following statistics are calculated:

Mean The average of all the values.

Minimum The smallest value.

Maximum The biggest value.

Std Dev (Standard Deviation) An expression of how widely spread the values are around the mean.

Class Width The x-axis distance between the left and right edges of each bar in the histogram.

Number of Classes The number of bars (including zero height bars) in the histograms.

Skewness Is the histogram symmetrical? If so, Skewness is zero. If the left hand tail is longer, skewness will be negative. If the right hand tail is longer, skewness will be positive. Where skewness exists, process capability indices are suspect. For process improvement, a good rule of thumb is to look at the long tail of your distribution; that is usually where quality problems lie.

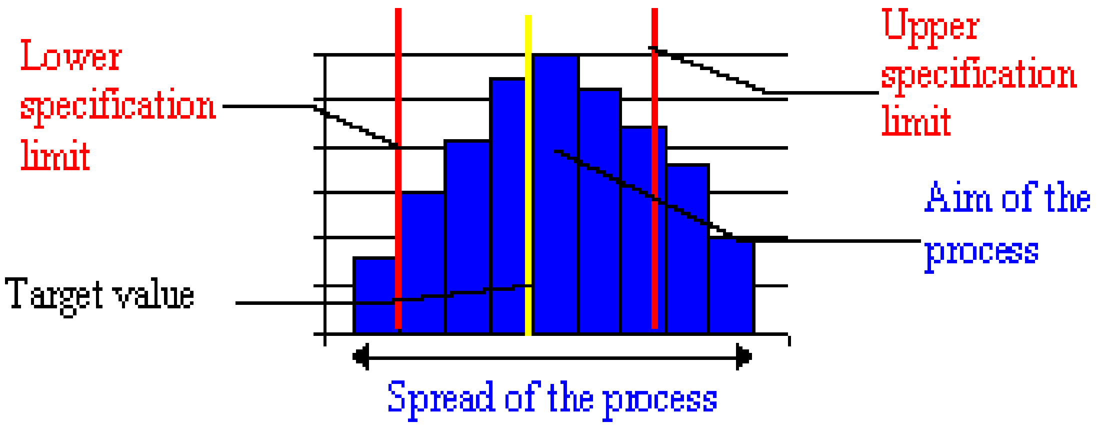
Kurtosis Kurtosis is a measure of the pointiness of a distribution. The standard normal curve has a kurtosis of zero. The Matterhorn, has negative kurtosis, while a flatter curve would have positive kurtosis. Positive kurtosis is usually more of a problem for quality control, since, with "big" tails, the process may well be wider than the spec limits.

Specification Limits and Batch Performance

Where relevant, you should display specification limits on your histograms. The specifications include a target value, an upper limit and a lower limit. For example, if Michael Jordan is shooting a basketball at a hoop, his target is the middle of the hoop. His spec limits are those points in the circle of the hoop that will just allow the ball to bounce through the basket. If the shot is outside spec limits, the ball doesn't go in.

[1] http://www.skymark.com/resources/tools/histograms.asp

When you overlay specification limits on a histogram, you can estimate how many items are being produced which do not meet specifications. This gives you an idea of batch performance, that is, of how the process performed during the period that you collected data. When you have added target, upper and lower limit lines, you can examine your histogram to see how your process is performing.



If the histogram shows that your process is wider than the specification limits, then it is not presently capable of meeting your specifications. This means the variation of the process should be reduced.

Also, if the process is not centered on the target value, it may need to be adjusted so that it can, on average, hit the target value. Sometimes, the distribution of a process could fit between the specification limits if it was centered, but spreads across one of the limits because it is not centered. Again, the process needs to be adjusted so that it can hit the target value most often.

Center of a Distribution

Processes have a target value, the value that the process should be producing, where most output of the process should fall. The center of the distribution in a histogram should, in most cases, fall on or near this target value. If it does not, the process will often need to be adjusted so that the center will hit the target value.

Spread of a Distribution

The spread, or width of a process is the distance between the minimum and maximum measured values. If the spread of the distribution is narrower than the specification limits, it is an indication of small variability in the process. This is almost always the goal, since consistency is important in most processes. If the distribution is wider than the specification limits the process has too much variability. The process is generating products that do not conform to specifications, i.e. junk.

Shape: Skewness and Kurtosis

A "normal" distribution of variation results in a specific bell-shaped curve, with the highest point in the middle and smoothly curving symmetrical slopes on both sides of center. The characteristics of the standard normal distribution are tabulated in most statistical reference works, allowing the relatively easy estimation of areas under the curve at any point.

Many distributions are non-normal. They may be skewed, or they may be flatter or more sharply peaked than the normal distribution.

A "skewed" distribution is one that is not symmetrical, but rather has a long tail in one direction. If the tail extends to the right, the curve is said to be right-skewed, or positively skewed. If the tail extends to the left, it is negatively skewed. Where skewness is present, attention should usually be focused on the tail, which could extend beyond the process specification limits, and where much of the potential for improvement generally lies.

Kurtosis is also a measure of the length of the tails of a distribution. For example, a symmetrical distribution with positive kurtosis indicates a greater than normal proportion of product in the tails. Negative kurtosis indicates shorter tails than a normal distribution would have.

Taken together, the values for process center, spread, skewness and kurtosis can tell you a great deal about your process. However, unless you have a solid statistics background, you will probably learn more from looking at the histogram itself than from looking at the statistics. Just remember that, where there is data in the tails near a specification limit, chances are that some non-conforming product is being made. If your process is actually making 5 bad parts in every thousand, and you are sampling 20 in every thousand, it will take some time before you find any out-of-spec parts. There are three things you should do:

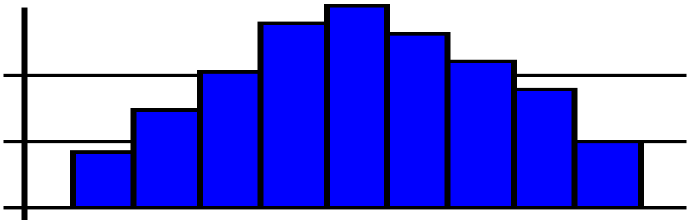
1. keep tracking data
2. get help in fitting a curve to your distribution
3. make sure your sampling plan is efficient.

Distributions you may encounter[2]

[2] **http://deming.eng.clemson.edu/pub/tutorials/qctools/histm.htm**

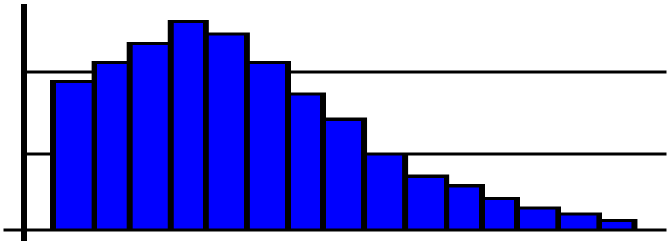
Normal

- Depicted by a bell-shaped curve
  - Most frequent measurement appears as center of distribution
  - Least frequent measurements taper gradually at both ends of distribution
- Indicates that a process is running normally (only common causes are present)



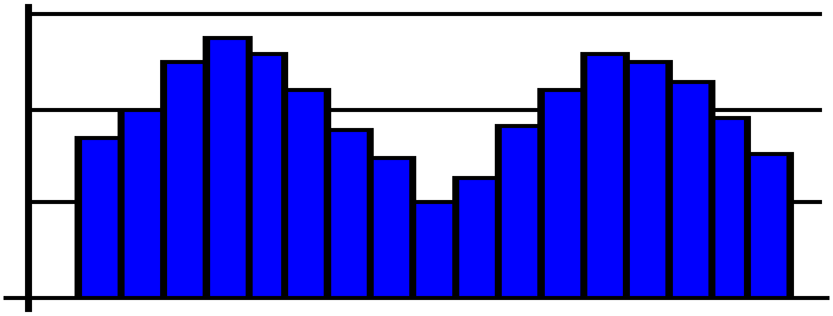
Skewed

- Appears as an uneven curve (with one tail longer than the other); values seem to taper to one side



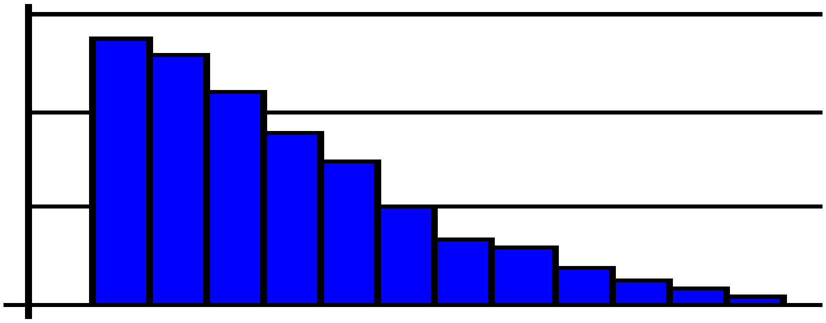
Bi-Modal (double-peaked)

- Distribution appears to have two peaks
- May indicate that data from more than one process are mixed together
  - Materials may have come from two separate vendors
  - Samples may have come from two separate machines
- A bi-modal curve often means that the data actually reflects two distinct processes with different centers. You will need to distinguish between the two processes to get a clear view of what is really happening in either individual process.

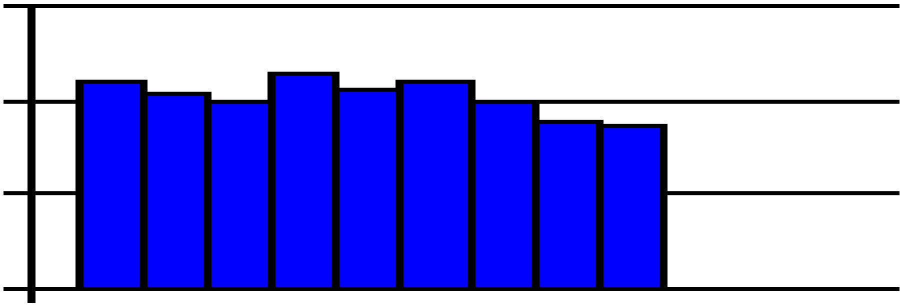


Cliff-Like

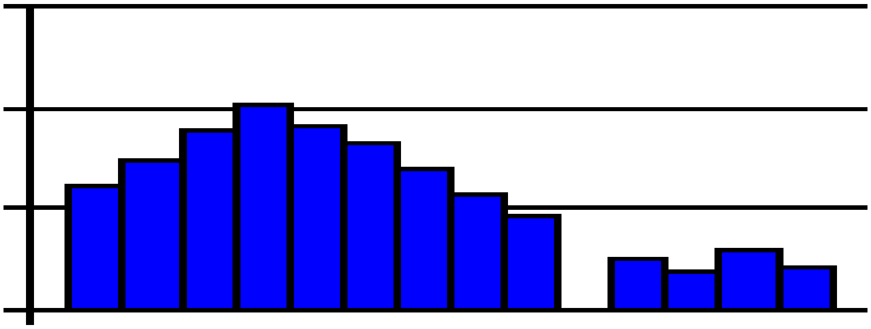
- Appears to end sharply or abruptly at one end
- Indicates possible sorting or inspection of non-conforming parts
- A truncated curve, with the peak at or near the edge while trailing gently off to the other side, often means that part of the distribution has been removed through screening, 100% inspection, or review. These efforts are usually costly and make good candidates for improvement efforts.



- A plateau-like curve often means that the process is ill-defined to those doing the work, which leaves everyone on their own. Since everyone handles the process differently, there are many different measurements with none standing out. The solution here is to clearly define an efficient process.



- Outliers in a histogram – bars that are removed from the others by at least the width of one bar – sometimes indicate that perhaps a separate process is included, but one that doesn't happen all the time. It may also indicate that special causes of variation are present in the process and should be investigated, though if the process is in control before the histogram is made as it should be, this latter option is unlikely.



Procedure for constructing a histogram

- Count the number of data points you have collected. (To produce a valid histogram, you should have at least 30 data points.)
- Determine the highest and lowest numbers from the collected data. Calculate the numerical difference between these two numbers. This will provide you with the "range".
- Divide the range into equally large "classes". These "classes" are actually the number of "bars" that will appear on your histogram. An easy way to determine the number of "classes" is to divide that range by 1, 2, 5, 10, 20, etc. You generally like to see anywhere from 6-15 "classes" on your chart.
- Next you will have to determine the width of each "class"/bar. In order to calculate the width, divide the "range" by the number of "classes". The width should have as many decimal points as the data points. Determine the lower and upper values for the individual "classes" by setting the smallest value of the data set as the lower value for the first "class". Find the upper value for this "class" by adding the "class" width (as determined above) to the lower value. The higher value of one "class" in turn becomes the lower value for the next "class". The lower value is always included in its "class" (that is, ≥ lower value), while the upper value belongs to the next class (that is, < upper value).
- To simplify the construction of the histogram, insert the data into a check sheet.
- Construct the histogram based on the check sheet. Mark the "classes" along the horizontal axis and the frequency along the vertical axis. Use vertical bars to indicate the distribution among "classes".

The following shows examples for each of the steps listing in constructing a histogram[3]:

1. Determine the range of the data by subtracting the smallest observed measurement from the largest and designate it as R.

   Example:
   Largest observed measurement = 1.1185 inches
   Smallest observed measurement = 1.1030 inches

   R = 1.1185 inches - 1.1030 inches =.0155 inch

2. Record the measurement unit (MU) used. This is usually controlled by the measuring instrument least count.

   Example: MU = .0001 inch

3. Determine the number of classes and the class width. The number of classes, k, should be no lower than six and no higher than fifteen for practical purposes. Trial and error may be done to achieve the best distribution for analysis.

   Example: k=8

4. Determine the class width (H) by dividing the range, R, by the preferred number of classes, k.

   Example: R/k = .0155/8 = .0019375 inch

The class width selected should be an odd-numbered multiple of the measurement unit, MU. This value should be close to the H value:
MU = .0001 inch
Class width = .0019 inch or .0021 inch

5. Establish the class midpoints and class limits. The first class midpoint should be located near the largest observed measurement. If possible, it should also be a convenient increment. Always make the class widths equal in size, and express the class limits in terms which are one-half unit beyond the accuracy of the original measurement unit. This avoids plotting an observed measurement on a class limit.

Example: First class midpoint = 1.1185 inches, and the class width is .0019 inch. Therefore, limits would be 1.1185 + or - .0019/2.

[3] **http://www.skymark.com/resources/tools/histograms.asp**

6. Determine the axes for the graph. The frequency scale on the vertical axis should slightly exceed the largest class frequency, and the measurement scale along the horizontal axis should be at regular intervals which are independent of the class width.
7. Draw the graph. Mark off the classes, and draw rectangles with heights corresponding to the measurement frequencies in that class.
8. Title the histogram. Give an overall title and identify each axis.

Example of a Histogram in Action[4]

Let us assume we have a performance indicator for Lost or Restricted Workdays due to occupational illness and injuries. Further, we are concerned because the past seven months have been above the baseline average (an indication of a significant increase), and we need to determine the source of this trend. The time period for histogram analyses will be the past seven months.

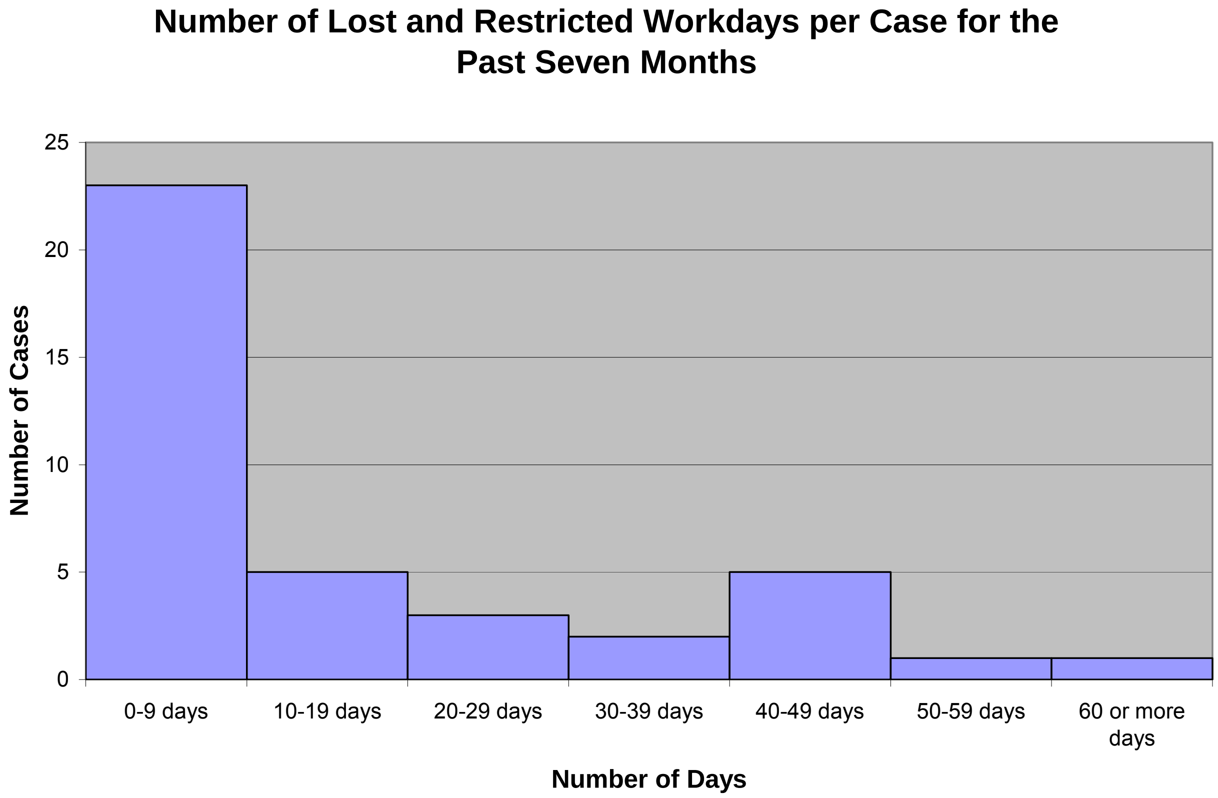
A histogram analysis could be used to look at the distribution of the number of days in each injury and illness case. The histogram will show if we have a problem with a large number of cases with a small number of days each, or a small number of cases with a large number of days each.

Steps in Making the Example Histogram

Step	Example
Define the data	The Number of Lost or Restricted Work Days per Case
Define the time period for the data	Past seven months of cases
Tabulate the data	List the number of days in each case: 47, 1, 55, 30, 1, 3, 7, 14, 7, 66, 34, 6, 10, 5, 12, 5, 3, 9, 18, 45, 5, 8, 44, 42, 46, 6, 4, 24, 24, 34, 11, 2, 3, 13, 5, 5, 3, 4, 4, 1
Determine the Range of the data (minimun value and maximum value).	The data ranges from a minimum of 1 day to a maximum of 66 days per case
Decide on the number of classes, and the width of each class (usually 6 - 15 equally spaced classes over the range of the entire data)	Use seven classes in ten day increments starting from zero (0-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60 or more). This gives a convenient grouping of the data in easy to work with increments, covering from 1 to 66.

[4] http://www.hanford,gov/safety/vpp/histo.htm

Count the number of items in each class	23 cases ranged from 0 to 9 days, 5 ranged from 10 to 19 days, 3 from 20 to 29 days, 2 from 30 to 39 days, 5 from 40 to 49 days, 1 from 50 to 59 days, and 1 involved more than 60 days.
Make a bar chart of the data using graph paper or a computer graphics routine	See example made using Excel spreadsheet below


Number of Lost and Restricted Workdays per Case for the Past Seven Months

There were seven equal width intervals chosen for the histogram. The histogram shows that most cases involve less than ten days, with a tapering off of cases in the higher value bins. The exception is 40 to 49 day bin, which is higher than the previous two bins. This may be the source of our increase and these five cases may need to be looked at in detail. A comparison with the shape of a histogram of cases prior to the increase may be worthwhile.

Some examples of histogram analyses use unequal width classes. This should be avoided if at all possible. Most observers expect the bins to be of equal widths, and one can distort the data presentation (and affect the conclusions drawn) by manipulating interval widths.

Other sources of information used for the content of this paper include:

http://www.isixsigma.com/library/content/c010527c.asp

Andersen, Bjorn and Tom Fagerhaug. Root Cause Analysis; Simplified Tools and Techniques. Milwaukee, WI; ASQ Quality Press, 2000.

Okes, Duke and Russ Westcott, eds. The Certified Quality Manager Handbook, Second Edition. Milwaukee, WI; ASQ Quality Press, 2001.